Simple Rules, Not So Simple

Thieme

Simple Rules, Not So Simple: The Use of International Ovarian Tumor Analysis (IOTA) Terminology and Simple Rules in Inexperienced Hands in a Prospective Multicenter Cohort Study

„Simple Rules“ – nicht so einfach: Anwendung der „International Ovarian Tumor Analysis“ (IOTA)-Terminologie und der „Simple Rules“ in unerfahrenen Händen in einer prospektiven multizentrischen Kohortenstudie

Authors

Evelyne Meys[1], Iris Rutten[1], Roy Kruitwagen[1], Brigitte Slangen[1], Sandrina Lambrechts[1], Helen Mertens[2], Ernst Nolting[3], Dieuwke Boskamp[4], Toon Van Gorp[1]

Affiliations

1 GROW – School for Oncology and Developmental Biology, Maastricht University Medical Centre (MUMC+), Maastricht, Netherlands
2 Board of Directors, Maastricht University Medical Centre (MUMC+), Maastricht, Netherlands (Formerly: Zuyderland hospital, Sittard)
3 Obststrics & Gynaecology, Sint Jans Gasthuis Weert, Netherlands
4 Obststrics & Gynaecology, VieCuri Medisch Centrum, Venlo, Netherlands

Key words

ultrasound, adnexal mass, IOTA simple rules, reproducibility, ovarian cancer

received 13.10.2016
accepted 29.04.2017

Bibliography

DOI https://doi.org/10.1055/s-0043-113819
Published online: August 23, 2017 | Ultraschall in Med 2017; 38: 633–641

ISSN 0172-4614

Correspondence

Evelyne Meys
Obststrics & Gynaecology, Maastricht Universitair Medisch Centrum+, P. Debyelaan 25, 6202 AZ Maastricht, Netherlands
Tel.: ++31/43/3 87 47 67
evelyne.meys@mumc.nl

ABSTRACT

Objectives To analyze how well untrained examiners – without experience in the use of International Ovarian Tumor Analysis (IOTA) terminology or simple ultrasound-based rules (simple rules) – are able to apply IOTA terminology and simple rules and to assess the level of agreement between non-experts and an expert.

Methods This prospective multicenter cohort study enrolled women with ovarian masses. Ultrasound was performed by non-expert examiners and an expert. Ultrasound features were recorded using IOTA nomenclature, and used for classifying the mass by simple rules. Interobserver agreement was evaluated with Fleiss' kappa and percentage agreement between observers.

Results 50 consecutive women were included. We observed 46 discrepancies in the description of ovarian masses when non-experts utilized IOTA terminology. Tumor type was misclassified often (n = 22), resulting in poor interobserver agreement between the non-experts and the expert (kappa = 0.39, 95 %-CI 0.244 – 0.529, percentage of agreement = 52.0 %). Misinterpretation of simple rules by non-experts was observed 57 times, resulting in an erroneous diagnosis in 15 patients (30 %). The agreement for classifying the mass as benign, malignant or inconclusive by simple rules was only moderate between the non-experts and the expert (kappa = 0.50, 95 %-CI 0.300 – 0.704, percentage of agreement = 70.0 %). The level of agreement for all 10 simple rules features varied greatly (kappa index range: -0.08 – 0.74, percentage of agreement 66 – 94 %).

Conclusion Although simple rules are useful to distinguish benign from malignant adnexal masses, they are not that simple for untrained examiners. Training with both IOTA terminology and simple rules is necessary before simple rules can be introduced into guidelines and daily clinical practice.

ZUSAMMENFASSUNG

Ziel Analyse, ob ungeübte Anwender – ohne Erfahrung mit der „International Ovarian Tumor Analysis“ (IOTA)- Terminologie oder einfache ultraschallbasierten Regeln („Simple Rules“) – in der Lage sind, IOTA-Kriterien und „Simple Rules“ anzuwenden. Auch wird der Grad der Übereinstimmung zwischen Nichtexperten und Experten bewertet.

Methoden Diese prospektive multizentrische Kohortenstudie nahm Frauen mit ovarialen Raumforderungen auf. Die Sonografie wurde von Nichtexperten und einem Experten durchgeführt. Ultraschall-Kriterien wurden mittels IOTA-Nomenklatur dokumentiert und dann für die Klassifizierung der Raumforderung mittels „Simple Rules“ verwendet. Die Intraobserver-Übereinstimmung zwischen den Beobachtern wurde durch Fleiss-Kappa und prozentualer Übereinstimmung bewertet.

Ergebnisse Eingeschlossen wurden 50 aufeinander folgende Frauen. Wir beobachteten 46 Diskrepanzen bei der Beschreibung der ovarialen Raumforderungen, wenn Nichtexperten die IOTA-Terminologie benutzten. Der Tumortyp wurde häufig falsch klassifiziert (n = 22), was zu einer schlechten Interobserver-Übereinstimmung zwischen Nichtexperten und Experten führte (Kappa = 0,39; 95 %-CI 0,244 – 0,529; prozentuale Übereinstimmung = 52,0 %). Eine Falschinterpretation der „Simple Rules“ durch Nichtexperten wurde 57 Mal beobachtet und führte bei 15 Patienten (30 %) zu einer Falschdiagnose. Die Übereinstimmung bei der Klassifizierung einer Raumforderung als gutartig, maligne oder nicht eindeutig durch die „Simple Rules“ war zwischen Nichtexperten und Experten nur mittelmäßig (Kappa = 0,50; 95 %-CI 0,300 – 0,704; prozentuale Übereinstimmung = 70,0 %). Der Grad der Übereinstimmung bei allen 10 „Simple Rules“-Kriterien variierte enorm (Kappa-Index-Bereich: -0,08 – 0,74; prozentuale Übereinstimmung 66 – 94 %).

Schlussfolgerung Obwohl die „Simple Rules“ nützlich sind, um benigne und maligne adnexale Raumforderungen zu unterscheiden, sind diese für ungeübte Untersucher nicht so einfach zu handhaben. Schulungen die sowohl IOTA-Terminologie als auch „Simple Rules“, zum Inhalt haben sind notwendig, noch ehe „Simple Rules“ in Leitlinien und den Praxisalltag Eingang finden.

Introduction

Ultrasound is an indispensable tool in the preoperative diagnosis of ovarian cancer. Correct characterization of an adnexal mass is important to ensure optimal management of the mass. In order to differentiate benign from malignant ovarian masses, many different ultrasound models and scoring systems have been developed over recent years. However, subjective assessment of ultrasound images by an expert examiner is considered the best way to classify these masses [1]. Nonetheless, it takes years of training and experience to become an expert. Therefore, other methods are needed to help less experienced ultrasonographers differentiate benign from malignant adnexal masses.

The International Ovarian Tumor Analysis (IOTA) Group developed ten clinically useful ultrasound rules to characterize ovarian masses as benign or malignant [2, 3]. This method is called simple ultrasound-based rules (simple rules) and contains five ultrasound features suggestive of a benign tumor and five features suggestive of a malignant tumor. The simple rules have been externally validated in several studies in which sensitivities of 73 – 100 % and specificities of 60 – 97.5 % have been reported [3 – 13]. However, in order to apply the simple rules, one needs to be familiar with IOTA terms and definitions as described in a consensus paper [14]. Thus far, the simple rules have only been validated by examiners trained in the use of IOTA terminology.

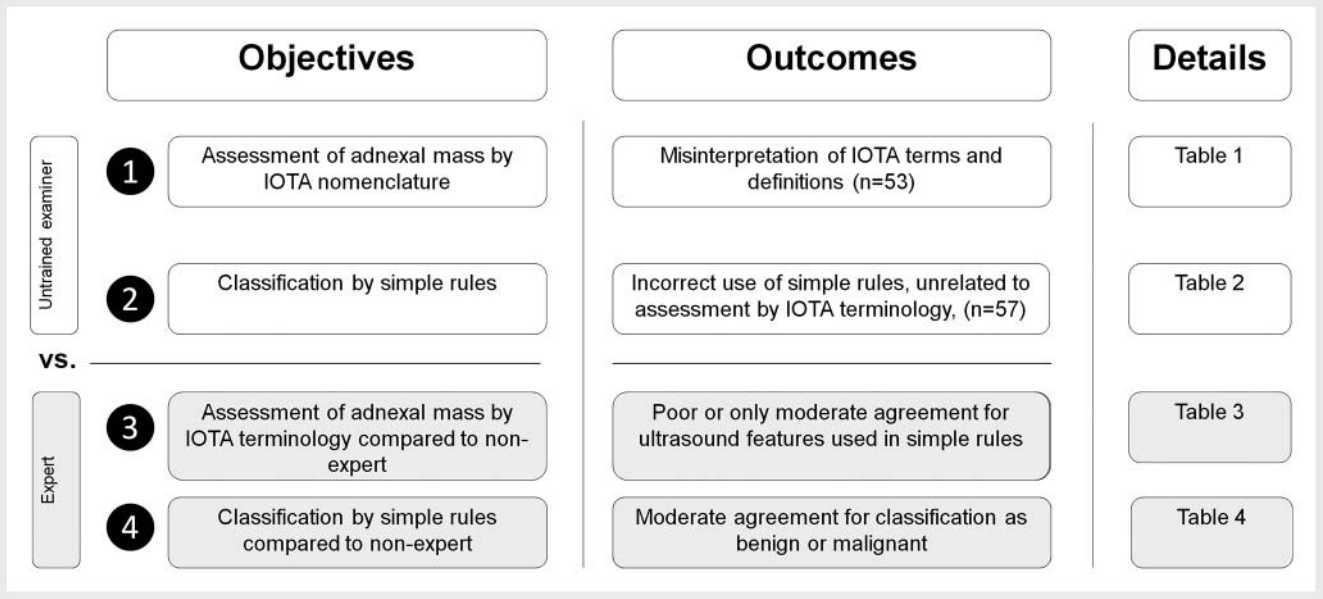
The aim of this study was to analyze the applicability of the simple rules when used by untrained examiners, i. e. with no experience in the use of either IOTA terminology or the simple rules. Therefore, four assessments were made (▶ **Fig. 1**):

1. How well is the IOTA terminology applied by untrained examiners?
2. How well are the simple rules applied by untrained examiners?
3. What is the level of agreement between untrained examiners and an expert examiner for utilization of the IOTA terminology?
4. What is the level of agreement between untrained examiners and an expert examiner for the classification of adnexal masses according to the IOTA simple rules?

Materials and methods

This was a prospective multicenter cohort study, called the SUBSONiC study (Simple Ultrasound-Based ruleS to differentiate OvariaN Cysts) [15]. Consecutive patients were recruited in a tertiary referral center – Maastricht University Medical Centre+ (MUMC+) – and three regional hospitals: Viecuri Venlo, Zuyderland hospital Sittard (formerly: Orbis medical centre Sittard), and St. Jans hospital Weert. Eligible patients had to be 18 years of age or older and diagnosed in one of the participating centers with a pelvic mass suspected to be of ovarian origin. Exclusion took place for: (a) pregnant patients; (b) patients with a prior bilateral oophorectomy in their history; (c) patients from whom sufficient data could not be retrieved; (d) patients who did not give or were incapable of giving informed consent; and (e) patients unable or unwilling to travel to the MUMC+ for a second ultrasound scan by an expert examiner (TvG).

Prior to the start of the study, theoretical training of approximately 2 hours was conducted for the untrained ultrasound examiners participating in the study. During this training, IOTA definitions of the ultrasound features adopted in the simple rules were explained by the expert examiner and some examples in which the definitions and simple rules were practiced were discussed.



▶ **Fig. 1** Study objectives and outcomes.

The study was approved by the local research ethics committees of all participating hospitals (NL44 181.068.13). All women included in the study gave written informed consent. STARD guidelines were followed for the conduct, analysis and reporting of our study [16].

Originally we anticipated performing a prospective multicenter diagnostic test accuracy study for the simple rules [15]. We prematurely stopped this study after an interim analysis of 50 patients. The results of the interim analysis are the subject of this article.

Ultrasound

All women underwent transvaginal, transrectal and/or transabdominal grayscale and color Doppler ultrasound. The first ultrasound scan was performed during the initial visit of the patient at the outpatient clinic by a non-expert examiner, i. e. a level-I or II examiner according to EFSUMB criteria (European Federation of Societies for Ultrasound in Medicine and Biology) [17]. A standardized approach was used and ultrasound features were recorded meticulously in a predefined data collection form using the nomenclature of the IOTA Group [14]. Further details regarding data collection can be found in supplementary file S1. After this assessment the ultrasonographer noted which of the simple rules were applicable and what the final diagnosis based on the simple rules was [2, 3]. The simple rules consist of ten ultrasound features: five features suggestive of a benign tumor (B-features) and five suggestive of a malignancy (M-features). The B-features are: unilocular cyst, the presence of solid components where the largest solid component has the largest diameter < 7 mm, the presence of acoustic shadows, smooth multilocular tumor with largest diameter < 100 mm, and no blood flow (color score 1). The M-features are: irregular solid tumor, presence of ascites, at least four papillary structures, irregular multilocular solid tumor with largest diameter ≥ 100 mm and very strong blood flow (color score 4).

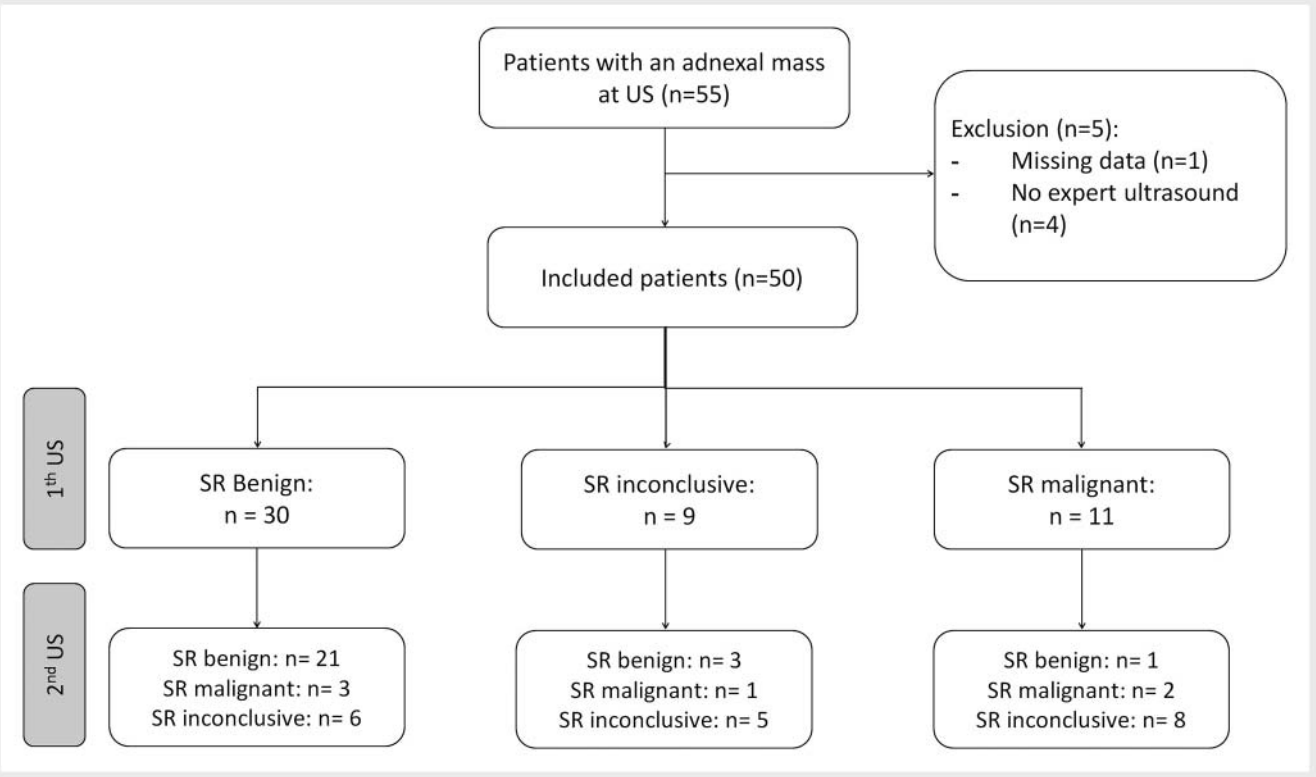
If one or more B-features are present in the absence of M-features, the tumor is diagnosed as benign. Vice versa, if one or more M-features are present in the absence of B-features, the mass is classified as malignant. In case both B- and M-features are present or if none of the ten features is present, the mass is classified as inconclusive.

Furthermore, we collected information regarding age, menopausal status, use of contraceptives (if any), parity, medical history, family history of breast or ovarian carcinoma, physical complaints and tumor markers.

Subsequently, subjects underwent a second ultrasound, performed by a single level-III examiner according to EFSUMB guidelines (TvG) using a Voluson E8 machine (GE Healthcare, Milwaukee, II, USA). This expert has almost 20 years of experience, has taken several courses from the IOTA group and is experienced in the use of the IOTA nomenclature and simple rules. The same ultrasound features as described in supplementary file S1 were assessed by this examiner, and a classification of each mass based on the simple rules was made. The descriptions of ultrasound features and outcome of the simple rules as evaluated by the expert examiner were used as the reference standard for our comparisons.

All data were entered into a clinical research form and later filed in a specially designed, secure data collection system (MACRO, Version 4.2.3.3850 InferMed Limited, London, UK). Since this ultrasound examination only took place after examination by a non-expert, level-I and II examiners were blinded to the results of the expert ultrasound. The expert examiner was not blinded to the results of the initial assessment.

In the course of the study, we had to conclude that (a) definitions used to describe the masses were applied incorrectly, and (b) the simple rules were not interpreted correctly. It was therefore decided to end the study. It is after all impossible to calculate test performance, if the test is not conducted well.



▶ **Fig. 2** Flowchart of patients included in the study. The first ultrasound was performed by a level – I or II examiner (non-expert) with no experience using IOTA terms and definitions, and was followed by a second ultrasound by an expert (level – III examiner). Abbreviations: US: ultrasound; SR: IOTA simple ultrasound-based rules.

Statistical analysis

Interobserver agreement was evaluated with Fleiss' kappa. The kappa value quantifies how much the observed agreement exceeds agreement by chance (values of 0.81 – 1.0 indicate very good agreement, 0.61 – 0.80 good agreement, 0.41 – 0.60 moderate agreement, 0.00 – 0.40 poor agreement) [18]. A value of zero indicates agreement equivalent to chance, while negative values indicate that the observed agreement is less than what is expected by chance. Since kappa values are affected by prevalence, and skewed data can result in low kappa values, we also calculated the absolute percentage agreement between the untrained examiners and the expert examiner [19].

In women with bilateral tumors, only the tumor with the most complex ultrasound morphology was included in the comparison. If both masses had the same morphology, the mass with the largest size was used for statistical analysis. Borderline tumors were classified as malignant.

All statistical analyses were conducted with IBM SPSS statistics version 20 (IBM Corp, Los Angeles, California, USA) and ReCal3, an online utility that computes interrater reliability coefficients [20].

Results

We enrolled 55 patients from September 2014 until September 2015 (▶ **Fig. 2**). Five patients were excluded: one due to missing data (incomplete ultrasound assessment because the study protocol was not followed completely) and four because no ultrasound by the expert examiner took place (one patient went elsewhere for further diagnosis and treatment and the others underwent surgery quickly leaving no time for the expert examiner to make an ultrasound). Ultimately, 50 patients were included in the study. The median age was 64.5 years (range: 27 – 91 years), and 14 (28 %) patients were premenopausal (▶ **Supplementary Table 1**, supplementary file S2). The first ultrasound examination was conducted by a resident (level-I) in 35 patients, and a gynecologist (level-II) in 15 patients. A total of 17 residents and 9 gynecologists, who examined between one and four patients each, participated in the study. All residents were in the third (n = 9), fourth (n = 6) or fifth (n = 2) year of their training. The group of level-II examiners consisted of staff gynecologists trained in ultrasound, but without special interest in ultrasound; 4 examiners were specialized in benign gynecology, two examiners specialized in gynecologic oncology and three examiners were general gynecologists with a special interest in oncology. The second ultrasound took place between 0 and 108 days (median: 7 days) after the first ultrasound.

The simple rules were applicable to 41 patients (82 %) when interpreted by non-experts and in 37 patients (74 %) when interpreted by the expert examiner.

Four assessments were made in accordance with the objectives (▶ **Fig. 1**). First, when evaluating the use of IOTA terminology by non-experts, we observed 46 discrepancies within the description of the ovarian mass that were mainly due to misinterpretation of

▶ **Table 1** Misinterpretation of IOTA terms and definitions by non-expert examiners for 50 patients.[1]

type of error	number of mistakes	
wrong tumor type		22 (48 %)
classified as unilocular instead of multilocular	1	
classified as unilocular instead of unilocular-solid	5	
classified as unilocular-solid instead of unilocular	1	
classified as unilocular-solid instead of solid	1	
classified as multilocular instead of unilocular	4	
classified as multilocular instead of multilocular-solid	9	
classified as multilocular-solid instead of multilocular	1	
measurement errors		20 (43 %)
ovarian mass not measured in 3 dimensions	7	
solid component not measured in 3 dimensions	8	
no color doppler assessment	3	
measurement of separate loci instead of total mass	2	
miscellaneous		4 (9 %)
incorrect interpretation of 'ascites'	2	
incorrect interpretation of 'regularity of the inner wall'	2	
	total	46

[1] Timmerman D, Valentin L, Bourne TH, Collins WP, Verrelst H, Vergote I, et al. Terms, definitions and measurements to describe the sonographic features of adnexal tumors: a consensus opinion from the International Ovarian Tumor Analysis (IOTA) Group. Ultrasound Obstet Gynecol. 2000;16(5):500 – 5

IOTA terms and definitions (▶ **Table 1**). In particular, tumor type was often misclassified (n = 22) and measurements of the mass were not performed in accordance with the IOTA guidelines (n = 20).

Second, the classification of adnexal masses by the simple rules applied by the untrained examiners was analyzed. Regardless of mistakes made in the description of the ovarian mass by IOTA nomenclature, we also observed misinterpretation of the simple rules themselves (i. e., wrong transfer from description of the mass to the simple rules). The simple rules were applied incorrectly 57 times by the non-experts (▶ **Table 2**). This incorrect use of the simple rules ultimately resulted in an erroneous diagnosis in 11 patients (22 %): 6 patients were diagnosed with a benign mass, while correct classification should have yielded an inconclusive result; 3 patients were diagnosed with a malignant mass, while correct classification should have led to an inconclusive result; and 2 patients in which the simple rules yielded an inconclusive result should have been diagnosed with a benign and malignant mass, respectively.

Third, when comparing the interpretation of IOTA terminology used by untrained examiners to the evaluation by the expert, we found frequent misclassification of tumor type by untrained examiners. This resulted in poor interobserver agreement regarding tumor type between the non-experts and the expert (kappa = 0.39, 95 % CI 0.244 – 0.529, percentage of agreement = 52.0 %) (▶ **Table 3**). The agreement for color Doppler score was poor as well (kappa = 0.19, 95 % CI 0.010 – 0.380, percentage of agreement = 46 %). The interobserver agreement for the ultrasound items included in the RMI was also just moderate (kappa = 0.60, 95 % CI 0.31 – 0.88, 81 % agreement) when compared to the expert's interpretation.

The fourth comparison, in which the interpretation of the simple rules applied by non-experts was compared to the interpretation of the expert, led to a different outcome in 15 patients (30 %). Therefore, the agreement for classifying the mass as benign, malignant or inconclusive by the simple rules was only moderate between the non-experts and the expert (kappa = 0.50, 95 % CI 0.300 to 0.704, percentage of agreement = 70.0 %). The level of agreement for the B/M-features of the simple rules varied greatly: from an observed agreement less than what is expected by chance to good agreement (▶ **Table 4**). When the group of non-experts was split up into only level-I or only level-II examiners, the agreement with the expert was good (kappa = 0.63, 95 % CI 0.388 – 0.869, percentage of agreement = 77.1 %) and poor (kappa = 0.15, 95 % CI -0.232 – 0.538, percentage of agreement 53.3 %), respectively. In only 3 cases the exact same simple rules were declared applicable when the interpretation of the non-expert was compared to that of the expert.

Discussion

This is the first study that investigated the interobserver agreement of the simple rules between non-experts and experts performed in a real-time setting (i. e., not based on video clips), in which non-experts had no previous knowledge of IOTA *terminology* and no experience with the simple rules [13, 21 – 24]. Moreover, the causes of the moderate agreement between non-experts and experts were also analyzed. We encountered two types of mistakes: IOTA definitions were applied incorrectly, and the simple rules were not interpreted correctly.

Regarding the first type of mistake, we observed misclassification of tumor type in 22 cases, with poor interobserver agreement regarding this descriptive item. In a recent study a kappa value of 0.70 was found for tumor type [25]. This related to agreement between two expert ultrasonographers. We do believe that the interobserver agreement in our population can increase – and mistakes can be prevented- by more comprehensive training. This is also demonstrated by two studies reporting the impediments ultrasound examiners may encounter when describing adnexal masses by IOTA nomenclature [26, 27]. Interobserver agreement regarding tumor type in non-experts improved substantially after a consensus meeting. The authors urged for more precise definitions of different descriptive items (e. g. papillary projections, solid components and Doppler color score). Moreover, the general opinion in the literature is that training regarding how to recog-

▶ **Table 2** Overview of the incorrect use of the IOTA simple rules by non-experts (independent of the wrong description of the mass using IOTA nomenclature) and explanation of mistakes regarding incorrect use of the simple rules.

	assigned incorrectly (total number assigned)	reason for noncompliance with description (number of times assigned incorrectly)	
B1	13 (17)	unilocular-solid tumor type	(11)
		multilocular type	(2)
B2	6 (7)	measurement of ≥ 7 mm	(4)
		unilocular tumor type	(1)
		solid tumor type	(1)
B3	0 (8)	–	–
B4	9 (15)	no regular tumor	(5)
		(multilocular-)solid tumor type	(2)
		unilocular tumor type	(1)
		largest diameter > 100 mm	(1)
B5	6 (25)	not applied, while doppler score was 1	(3)
		at color doppler score 2	(2)
		only doppler score for one of the adnexa	(1)
M1	12 (16)	multilocular-solid tumor type	(8)
		unilocular-solid tumor type	(2)
		no assessment of inner wall	(1)
		multilocular tumor type	(1)
M2	1 (6)	not applied, while ascites was described	(1)
M3	2 (6)	no papillary projections	(2)
M4	7 (12)	solid tumor type	(3)
		multilocular tumor type	(2)
		regular tumor	(1)
		no assessment of inner wall	(1)
M5	1 (2)	color doppler score 3	(1)

nize ultrasound features and subsequently describe adnexal masses using IOTA terminology should be provided in order to achieve higher interobserver agreement and better diagnostic performance [11, 13, 21, 25, 28 – 30].

In other studies investigating interobserver agreement for ultrasound features, stored images or video clips instead of real-time ultrasound examinations were used [21, 31]. In these studies the first ultrasound scan is usually made by an expert providing clear images. In the present study the first ultrasound scan was not made by an expert, but rather by a level-I/II examiner, in accordance with normal clinical practice. Interestingly, agreement with the expert was higher for level-I examiners than for level-II examiners, which could be explained by the additional education given to this group as part of their residency. This again stresses the importance of IOTA terminology training. Also, some studies demonstrate a small decrease in the ultrasound learning curves in the advanced stages of experience [32], perhaps because difficult-to-diagnose patients are usually seen by examiners with more experience. However, no such selection was made in our study.

The second category of mistakes was incorrect transfer from the description of the ovarian mass to the appropriate rules of the simple rules. For example, the mass was correctly described as multilocular solid with an irregular cyst wall, but then rule M1 (irregular solid tumor) was incorrectly applied. This is not the first study observing the incorrect application of the simple rules. Alcazar et al. found that 12 % of malignant masses were miscategorized as benign by non-expert examiners applying the simple rules [6]. Also, Knafel et al. reported that as much as 50 % of their inconclusive results (20/40) were in fact due to misclassification [11].

Nonetheless, most studies conducted up until now have concluded that the simple rules are easy to use in routine clinical practice. The explanation for these contrasting findings in our study is twofold. First, ultrasonographers in other studies were familiar with the nomenclature as described by the IOTA group [14]. This was not the case in our study. It is essential to comprehend all IOTA definitions in order to be able to apply the simple rules. Inter-center differences were observed in a multicenter IOTA study, which – according to the authors – could be due to the dif-

▶ **Table 3** Interobserver variability for observers with different levels of expertise (non-experts vs. an expert) for ultrasound features used to describe an adnexal mass by IOTA terminology.

	agreement	kappa value (95 % CI)
tumor type (unilocular/unilocular-solid/multilocular/multilocular-solid/solid)	52 % (26/50)	0.39 (0.244 – 0.529)
number of loculations (0/1/2/3/4/5 – 10/> 10)	52 % (26/50)	0.37 (0.231 – 0.501)
papillary projections (0/1/2/3/> 3)	74 % (37/50)	0.13 (–0.059 – 0.327)
acoustic shadow (yes/no)	74 % (37/50)	0.34 (0.063 – 0.618)
inner cyst wall (regular/irregular/unable to measure)	66 % (33/50)	0.41 (0.179 – 0.633)
septations (≤ 3 mm/ ≥ 3 mm/no septations)	66 % (33/50)	0.49 (0.294 – 0.678)
color doppler score (1: no blood flow/2:minimal blood flow/3: moderate blood flow/4: intense blood flow)	46 % (23/50)	0.19 (0.010 – 0.380)
ascites (yes/no)	88 % (44/50)	0.55 (0.276 – 0.831)
metastasis (yes/no)	82 % (41/50)	0.08 (–0.197 – 0.358)

ference in the examiners' use of IOTA terms [8]. In the Netherlands IOTA terminology is not used on a regular basis, which also made it more difficult for the level-I/II examiners to interpret the simple rules in the present study.

Second, contrary to our study, many other studies were conducted by ultrasonographers that have some experience with gynecologic ultrasound. In the Netherlands the emphasis for training in ultrasound is on obstetric ultrasound. Therefore, residents and gynecologists are less experienced in gynecologic ultrasound than ultrasonographers who have undergone formal ultrasound training. However, lacking training in gynecologic ultrasound is not an issue only in the Netherlands [33]. When an ultrasonographer is experienced in gynecologic ultrasound, some characteristic features of the mass can be anticipated, as is also stated by Tantipalakorn et al. [9]. For example, in the case of a dermoid cyst, an experienced ultrasonographer will easily recognize the echogenic interior as cyst content rather than solid tissue. Thus, the level of experience of the ultrasonographer could bias the interpretation of the simple rules, leading to better results in studies where the simple rules were validated by examiners more experienced in gynecologic ultrasound.

Another factor that might have contributed to the simple rules being applied incorrectly is the unusual distribution of tumor types in a relative small study population. Three patients suffered from rare benign tumors (struma ovarii, an atypical presentation of a myoma and a benign cyst from an origin unidentified even by pathology) and 5 patients had borderline tumors that are known to be difficult to diagnose [34]. This could also explain why the number of inconclusive results was higher than usual (26 % vs. 19 % in other studies) when the simple rules were applied by the expert examiner [1].

Despite the relatively small sample size, our study has the advantage of a prospective design and was conducted in both oncology and non-oncology centers and by non-experienced and experienced ultrasonographers, which represents day-to-day clinical practice. Since we included a consecutive series of patients, employed deliberate exclusion criteria, and conducted the study in various centers, we think this provides a random sample which is generalizable to the rest of the population.

The number of patients is rather small to perform statistically significant calculations on test performance. Furthermore, histologic confirmation of the diagnosis was available in only 37 patients: 23 masses were benign and 14 malignant (including 5 borderline tumors) (▶ **Supplementary Table 2**, supplementary file S3). However, had we enrolled more patients this would not have given an accurate account of test performance of the simple rules, since IOTA terms and definitions and the simple rules themselves were not applied correctly by the non-experts.

Interpretative errors have not been described previously in the literature concerning gynecological ultrasonography. However, these errors have been studied in other fields, such as in emergency ultrasound [32, 35 – 38]. Not surprisingly, the number of errors decreased as the sonographers gained experience. It is debatable how much experience is required to adequately perform gynecologic ultrasound with an acceptable degree of error. In one study high diagnostic accuracy for the differentiation of adnexal masses by ultrasound was achieved after analysis of 200 cases [39]. Other studies state that trainees struggle to achieve minimal ultrasound competences and as a consequence, some may never acquire the

▶ **Table 4** Interobserver variability for observers with different levels of expertise (non-experts vs. an expert) for all ultrasound (B/M) features included in the IOTA simple rules expressed as Fleiss' kappa index and percentage of agreement.

	agreement	kappa value (95 % CI)
B1	76 %	0.41 (0.13 – 0.68)
B2	86 %	–0.08 (–0.35 – 0.20)
B3	76 %	0.34 (0.06 – 0.62)
B4	74 %	0.22 (–0.06 – 0.49)
B5	66 %	0.30 (0.02 – 0.57)
M1	70 %	0.03 (–0.25 – 0.30)
M2	94 %	0.74 (0.46 – 1.0)
M3	88 %	0.19 (–0.09 – 0.46)
M4	88 %	0.63 (0.35 – 0.90)
M5	90 %	–0.05 (–0.33 – 0.22)

basic skills and knowledge needed for independent practice [40, 41].

In conclusion, we believe that ultrasound examiners should be aware of the poorer performance of the simple rules if the performing clinician has only little knowledge of the IOTA terms and definitions. In our study a 2-hour training session was not enough to fully comprehend the IOTA nomenclature and therefore correctly apply the simple rules. Consequently, we believe further training with both IOTA terminology and the simple rules is necessary before the simple rules can be introduced into daily clinical practice. The introduction of the simple rules in national guidelines should go hand in hand with national training programs or courses. After all, the simple rules are just not that simple.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

[1] Meys EMJ, Kaijser J, Kruitwagen RF et al. Subjective assessment versus ultrasound based models to diagnose ovarian cancer: a systematic review and meta-analysis. Eur J Cancer 2016; 58: 17 – 29

[2] Timmerman D, Testa AC, Bourne T et al. Simple ultrasound-based rules for the diagnosis of ovarian cancer. Ultrasound Obstet Gynecol 2008; 31: 681 – 690

[3] Timmerman D, Ameye L, Fischerova D et al. Simple ultrasound rules to distinguish between benign and malignant adnexal masses before surgery: prospective validation by IOTA group. BMJ 2010; 341: c6839

[4] Fathallah K, Huchon C, Bats AS et al. [External validation of simple ultrasound rules of Timmerman on 122 ovarian tumors]. Gynecol Obstet Fertil 2011; 39: 477 – 481

[5] Hartman CA, Juliato CR, Sarian LO et al. Ultrasound criteria and CA 125 as predictive variables of ovarian cancer in women with adnexal tumors. Ultrasound Obstet Gynecol 2012; 40: 360 – 366

[6] Alcazar JL, Pascual MA, Olartecoechea B et al. IOTA simple rules for discriminating between benign and malignant adnexal masses: prospective external validation. Ultrasound Obstet Gynecol 2013; 42: 467 – 471

[7] Nunes N, Ambler G, Foo X et al. Use of IOTA simple rules for diagnosis of ovarian cancer: meta-analysis. Ultrasound Obstet Gynecol 2014; 44: 503 – 514

[8] Testa A, Kaijser J, Wynants L et al. Strategies to diagnose ovarian cancer: new evidence from phase 3 of the multicentre international IOTA study. Br J Cancer 2014; 111: 680 – 688

[9] Tantipalakorn C, Wanapirak C, Khunamornpong S et al. IOTA simple rules in differentiating between benign and malignant ovarian tumors. Asian Pac J Cancer Prev 2014; 15: 5123 – 5126

[10] Ruiz de Gauna B, Rodriguez D, Olartecoechea B et al. Diagnostic performance of IOTA simple rules for adnexal masses classification: a comparison between two centers with different ovarian cancer prevalence. Eur J Obstet Gynecol Reprod Biol 2015; 191: 10 – 14

[11] Knafel A, Banas T, Nocun A et al. The Prospective External Validation of International Ovarian Tumor Analysis (IOTA) Simple Rules in the Hands of Level I and II Examiners. Ultraschall in Med 2016; 37(5): 516 – 523

[12] Silvestre L, Martins WP, Candido-Dos-Reis FJ. Limitations of three-dimensional power Doppler angiography in preoperative evaluation of ovarian tumors. J Ovarian Res 2015; 8: 47

[13] Tinnangwattana D, Vichak-Ururote L, Tontivuthikul P et al. IOTA Simple Rules in Differentiating between Benign and Malignant Adnexal Masses by Non-expert Examiners. Asian Pac J Cancer Prev 2015; 16: 3835 – 3838

[14] Timmerman D, Valentin L, Bourne TH et al. Terms, definitions and measurements to describe the sonographic features of adnexal tumors: a consensus opinion from the International Ovarian Tumor Analysis (IOTA) Group. Ultrasound Obstet Gynecol 2000; 16: 500 – 505

[15] Meys EM, Rutten IJ, Kruitwagen RF et al. Investigating the performance and cost-effectiveness of the simple ultrasound-based rules compared to the risk of malignancy index in the diagnosis of ovarian cancer (SUBSONiC-study): protocol of a prospective multicenter cohort study in the Netherlands. BMC Cancer 2015; 15: 482

[16] Bossuyt PM, Reitsma JB, Bruns DE et al. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. Ann Intern Med 2003; 138: 40 – 44

[17] European Federation of Societies for Ultrasound in Medicine, Biology, Education, Committee PS. Minimum training recommendations for the practice of medical ultrasound. Ultraschall in Med 2006; 27: 79 – 105

[18] Brennan P, Silman A. Statistical methods for assessing observer variability in clinical measures. BMJ 1992; 304: 1491 – 1494

[19] Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol 1990; 43: 543 – 549

[20] Freelon DG. ReCal: Intercoder reliability calculation as a web service. International Journal of Internet Science 2010; 5: 20 – 23

[21] Ruiz de Gauna B, Sanchez P, Pineda L et al. Interobserver agreement in describing adnexal masses using the International Ovarian Tumor Analysis simple rules in a real-time setting and using three-dimensional ultrasound volumes and digital clips. Ultrasound Obstet Gynecol 2014; 44: 95 – 99

[22] Sayasneh A, Kaijser J, Preisler J et al. A multicenter prospective external validation of the diagnostic performance of IOTA simple descriptors and rules to characterize ovarian masses. Gynecol Oncol 2013; 130: 140 – 146

[23] Sayasneh A, Wynants L, Preisler J et al. Multicentre external validation of IOTA prediction models and RMI by operators with varied training. Br J Cancer 2013; 108: 2448 – 2454

[24] Alcazar JL, Pascual MA, Graupera B et al. External validation of IOTA simple descriptors and simple rules for classifying adnexal masses. Ultrasound Obstet Gynecol 2016; 48: 397 – 402

[25] Sladkevicius P, Valentin L. Interobserver agreement in describing the ultrasound appearance of adnexal masses and in calculating the risk of malignancy using logistic regression models. Clin Cancer Res 2015; 21: 594 – 601

[26] Zannoni L, Savelli L, Jokubkiene L et al. Intra- and interobserver agreement with regard to describing adnexal masses using International Ovarian Tumor Analysis terminology: reproducibility study involving seven observers. Ultrasound Obstet Gynecol 2014; 44: 100 – 108

[27] Zannoni L, Savelli L, Jokubkiene L et al. Intra- and interobserver reproducibility of assessment of Doppler ultrasound findings in adnexal masses. Ultrasound Obstet Gynecol 2013; 42: 93 – 101

[28] Peces RamaA, Llanos LlanosMC, Sanchez FerrerML et al. Simple descriptors and simple rules of the International Ovarian Tumor Analysis (IOTA) Group: a prospective study of combined use for the description of adnexal masses. Eur J Obstet Gynecol Reprod Biol 2015; 195: 7 – 11

[29] Van Holsbeke C, Daemen A, Yazbek J et al. Ultrasound methods to distinguish between malignant and benign adnexal masses in the hands of examiners with different levels of experience. Ultrasound Obstet Gynecol 2009; 34: 454 – 461

[30] Van Calster B, Van Hoorde K, Froyman W et al. Practical guidance for applying the ADNEX model from the IOTA group to discriminate between different subtypes of adnexal tumors. Facts Views Vis Obgyn 2015; 7: 32 – 41

[31] Guerriero S, Saba L, Ajossa S et al. Assessing the reproducibility of the IOTA simple ultrasound rules for classifying adnexal masses as benign or malignant using stored 3D volumes. Eur J Obstet Gynecol Reprod Biol 2013; 171: 157 – 160

[32] Blehar DJ, Barton B, Gaspari RJ. Learning curves in emergency ultrasound education. Acad Emerg Med 2015; 22: 574 – 582

[33] Green J, Kahan M, Wong S. Obstetric and Gynecologic Resident Ultrasound Education Project: Is the Current Level of Gynecologic Ultrasound Training in Canada Meeting the Needs of Residents and Faculty? J Ultrasound Med 2015; 34: 1583 – 1589

[34] Fischerova D, Zikan M, Dundr P et al. Diagnosis, treatment, and follow-up of borderline ovarian tumors. Oncologist 2012; 17: 1515 – 1533

[35] Gaspari RJ, Dickman E, Blehar D. Learning curve of bedside ultrasound of the gallbladder. J Emerg Med 2009; 37: 51 – 56

[36] Jang T, Aubin C, Naunheim R. Minimum training for right upper quadrant ultrasonography. Am J Emerg Med 2004; 22: 439 – 443

[37] Heegeman DJ, Kieke B Jr. Learning curves, credentialing, and the need for ultrasound fellowships. Acad Emerg Med 2003; 10: 404 – 405

[38] Kendall JL, Shimp RJ. Performance and interpretation of focused right upper quadrant ultrasound by emergency physicians. J Emerg Med 2001; 21: 7 – 13

[39] Alcazar JL, Diaz L, Florez P et al. Intensive training program for ultrasound diagnosis of adnexal masses: protocol and preliminary results. Ultrasound Obstet Gynecol 2013; 42: 218 – 223

[40] Tolsgaard MG, Rasmussen MB, Tappert C et al. Which factors are associated with trainees' confidence in performing obstetric and gynecological ultrasound examinations? Ultrasound Obstet Gynecol 2014; 43: 444 – 451

[41] Patel H, Chandrasekaran D, Myriokefalitaki E et al. The Role of Ultrasound Simulation in Obstetrics and Gynecology Training: A UK Trainees' Perspective. Simul Healthc 2016; 11: 340 – 344